Temporal fine structure influences voicing confusions for consonant identification in multi-talker babble

Vibha Viswanathan[1,*], Barbara G. Shinn-Cunningham[3], and Michael G. Heinz[1,2]

[1] *Weldon School of Biomedical Engineering, Purdue University, West Lafayette, Indiana, United States*
[2] *Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana, United States*
[3] *Neuroscience Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States*
[*] *Correspondence: viswanav@purdue.edu*

Abstract

To understand the mechanisms of speech perception in everyday listening environments, it is important to elucidate the relative contributions of different acoustics cues in transmitting phonetic content. Previous studies suggest that the energy envelopes of speech convey most speech content, while the temporal fine structure (TFS) can aid in segregating target speech from background noise. Despite the vast literature on TFS and speech intelligibility, the role of TFS in conveying additional speech content over what envelopes convey in complex acoustic scenes is poorly understood. The present study addresses this question using online psychophysical experiments to measure consonant identification in multi-talker babble for intelligibility-matched intact and 64-channel envelope-vocoded stimuli. Consonant confusion patterns revealed that listeners had a greater tendency in the vocoded (versus intact) condition to be biased towards reporting that they heard an unvoiced consonant, despite envelope and place cues being largely preserved. This result was replicated when babble instances were varied across independent experiments, suggesting that TFS conveys important voicing cues over what envelopes convey in multi-talker babble, a masker that is ubiquitous in everyday environments. This finding has implications for assistive listening devices that do not currently provide TFS cues, such as cochlear implants.

Keywords: consonant confusions, speech intelligibility, temporal coding, cocktail-party listening, cochlear implants

1 Introduction

Any acoustic signal can be decomposed into a slowly varying amplitude envelope, or temporal modulation, and a fast-varying temporal fine structure (TFS) (Hilbert, 1906). The cochlea decomposes sound input into a multi-channel representation organized by frequency, where each channel encodes the signal content in a relatively narrow band of frequencies around a different carrier frequency. The envelope and TFS information in each channel are then conveyed to the central nervous system through the ascending auditory pathway (Johnson, 1980; Joris and Yin, 1992). Elucidating the relative contributions of envelope and TFS cues to speech perception in everyday listening environments is important not just from a basic science perspective, but also for translation to clinical technologies such as cochlear implants, which do not currently provide TFS information.

Psychophysical studies suggest that speech content in quiet is largely conveyed by envelopes (Shannon et al., 1995; Smith et al., 2002). Psychophysical (Bacon and Grantham, 1989; Stone and

Moore, 2014), modeling (Dubbelboer and Houtgast, 2008; Relaño-Iborra et al., 2016), and electroencephalography (EEG) (Viswanathan et al., 2021) studies support the theory that in the presence of background noise, modulation masking of envelopes of target speech by distracting masker envelopes predicts speech intelligibility across diverse listening conditions. However, in addition to this contribution of envelopes to intelligibility, TFS may also play a role, especially in noisy listening environments (Lorenzi et al., 2006; Hopkins and Moore, 2010).

Psychophysical studies suggest that cues conveyed by TFS (e.g., pitch) (Smith et al., 2002) critically support object formation, perceptual scene segregation, and selective attention (Darwin, 1997; Shinn-Cunningham, 2008; Oxenham and Simonson, 2009). Moreover, using EEG, Viswanathan et al. (2021) showed that TFS cues influence target-speech envelope coding in the brain, which in turn predicts intelligibility across a range of backgrounds and distortions. Despite the extensive prior literature on TFS and speech intelligibility, the role of TFS in conveying additional speech content over what envelopes convey in complex listening conditions, and beyond its role in supporting scene segregation, is poorly understood.

Previous behavioral studies that used TFS-vocoded speech (e.g., Sheft et al., 2008) found that TFS can convey certain phonetic features with relatively high levels of information reception. However, they did not examine whether or not TFS conveyed important phonetic content over and above the information conveyed by envelopes (Smith et al., 2002). That is, while they examined the role of TFS when envelope cues are removed, they could not address the question of whether or not TFS cues are used in intact speech that has preserved envelope cues. Another limitation of previous studies that investigated the role of TFS in conveying speech content is that they used conditions that were not ecologically realistic. While some used speech in quiet (Rosen, 1992; Sheft et al., 2008), others presented speech in stationary noise (Gnansia et al., 2009; Swaminathan and Heinz, 2012). Ecologically relevant maskers such as multi-talker babble have not been utilized to study this problem.

The present study addresses these gaps using careful envelope-vocoding experiments designed to probe directly the role of TFS in conveying consonant information in realistic listening environments. Here, we used multi-talker babble as an ecologically relevant masker. We analyzed consonant confusion patterns (Miller and Nicely, 1955), grouping consonants into the categories based upon the features of voicing, place of articulation (POA), and manner of articulation (MOA). We then examined whether confusion patterns are altered between intelligibility-matched intact and 64-channel envelope-vocoded stimuli for consonants presented in multi-talker babble and separately in quiet (as a control). 64-channel envelope vocoding largely preserves cochlear-level envelopes (Viswanathan et al., 2021), allowing us to study the role of TFS in conveying speech content over and above what is conveyed by envelopes. Since TFS plays a role in segregation, vocoding at the same signal-to-noise ratio (SNR) as intact stimuli produces considerably lower intelligibility. Here we mitigate this intelligibility drop by providing a higher SNR for vocoded stimuli so that overall intelligibility is matched between intact and vocoded conditions. This allows us to fairly compare confusion patterns across the two conditions to investigate the relationship between TFS and speech content. Finally, given that consonants are transient sounds, we also examined whether effects were robust to changes in the local statistics of the masker; accordingly, we tested whether results were replicated when the specific instances (i.e., examples or realizations) of multi-talker babble varied across experiments.

We hypothesized that TFS does not convey speech content over and above what is conveyed by envelopes. As a result, we expected that once we matched intelligibility across conditions, confusion patterns would be the same for intact and envelope-vocoded stimuli corresponding to speech in (i) babble, and (ii) quiet. Below, we describe the experiments we used to test this hypothesis and our results as well as their implications.

2 Materials and Methods

2.1 Stimulus generation

20 consonants from the STeVI corpus (Sensimetrics Corporation) were chosen for the study. The consonants were /b/, /ʧ/, /d/, /ð/, /f/, /g/, /ʤ/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /ʃ/, /t/, /θ/, /v/, /z/, and /ʒ/. The consonants were presented in CV (consonant-vowel) context, where the vowel was always /a/. Note that although consonant confusions differ depending on context (Dubno and Levitt, 1981), the particular choice of context (e.g., the specific vowel used, and whether the consonant occurs before or after the vowel) used here does not matter since we are only concerned with the effects of vocoding in this study, and the context was fixed across these conditions. Each consonant was spoken by two female and two male talkers (to reflect real-life talker variability). The CV utterances were embedded in the carrier phrase: "You will mark /CV/ please" (i.e., in natural running speech). Stimuli were created for five different experimental conditions as described below:

1. **Speech in Babble (SiB):** Speech mixed in spectrally matched four-talker babble at -8 dB SNR. Here, the long-term spectrum of the target speech (including the carrier phrase) was matched with the average (across instances) long-term spectrum of four-talker babble.

2. **Vocoded Speech in Babble (Vocoded SiB):** SiB at 0 dB SNR subjected to 64-channel envelope vocoding. The vocoding process retained the peripheral envelopes, but replaced the stimulus fine structure with a noise carrier, in accordance with the procedure described in Qin and Oxenham (2003). The bandwidths used during vocoding were the same as those of cochlear filters of normal-hearing subjects (Glasberg and Moore, 1990).

3. **Speech in Quiet (SiQuiet):** Speech in quiet was used as a control condition.

4. **Vocoded Speech in Quiet (Vocoded SiQuiet):** SiQuiet subjected to 64-channel envelope vocoding (using the same procedure as for Vocoded SiB) was used to examine whether fine structure conveys speech content over and above what is conveyed by envelopes in quiet.

5. **Speech in Speech-shaped Stationary Noise (SiSSN):** Speech mixed in spectrally matched stationary gaussian noise, i.e., speech-shaped stationary noise, at -8 dB SNR. Here, the long-term spectra of the target speech (including the carrier phrase) and that of stationary noise were matched with the average (across instances) long-term spectrum of four-talker babble. The SiSSN condition was used for online data quality checking, given that lab-based confusion data are available in this condition (Phatak and Allen, 2007).

To create each SiB stimulus, a babble sample was randomly selected from a list comprising 72 different four-talker babble maskers obtained from the QuickSIN corpus (Killion et al., 2004). Similarly, a different realization of stationary noise was used for each SiSSN stimulus.

Prior to the main consonant identification study, an offline behavioral pilot study (with three subjects who did not participate in the actual online experiments) was used to determine the SNRs for the different experimental conditions. The SNRs for the intact and vocoded SiB conditions were chosen to correspond to an intelligibility value of roughly 60%, so that a sufficient number of confusions would be obtained for data analysis.

To verify that our vocoding procedure did not significantly change envelopes at the cochlear level, we extracted the envelopes at the output of 128 filters (equally spaced on an ERB scale and with normal cochlear bandwidths; Glasberg and Moore, 1990) both before and after vocoding for SiQuiet and SiB at 0 dB SNR, and for each of the different consonants and talkers in our study.

Note that the use of 128 filters allowed us to compare envelopes at both on-band filters (i.e., filters whose center frequencies matched with the sub-bands of the vocoder), and off-band filters (i.e., filters whose center frequencies were halfway between adjacent vocoder sub-bands on the ERB scale). The average correlation coefficient between envelopes before and after vocoding (across the different stimuli and cochlear filters, and after adjusting for any vocoder group delays) is about 0.9 (Fig. 1). This suggests that our 64-channel envelope-vocoding procedure leaves the within-band cochlear-level envelopes largely intact. This high-resolution vocoding allowed us to unambiguously attribute vocoding effects to TFS cues rather than any spurious envelopes (not present in the original stimuli) that are introduced within individual frequency bands during cochlear filtering of the noise carrier used in vocoding when low-resolution vocoding is performed (Gilbert and Lorenzi, 2006; Swaminathan and Heinz, 2012; Viswanathan et al., 2021).

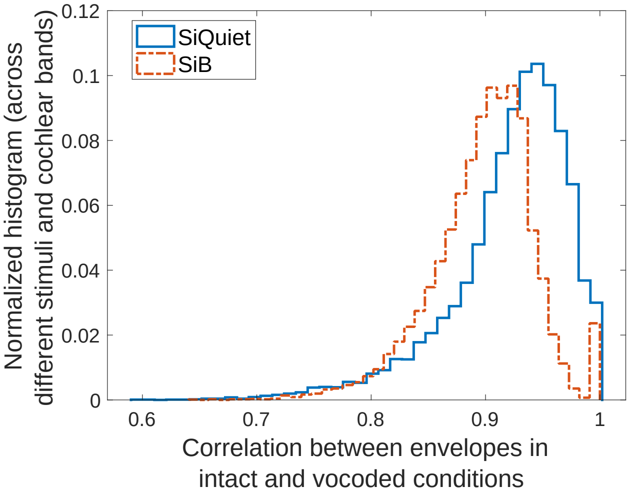

Figure 1. 64-channel envelope vocoding largely preserves the envelopes within individual cochlear bands. Shown are the normalized histogram of the group-delay-adjusted correlation between the envelope in intact speech in quiet (SiQuiet) and 64-channel vocoded SiQuiet (i), and that in intact speech in babble (SiB) and 64-channel vocoded SiB (ii). The histograms are across the different consonants and talkers in our study, as well as across 128 different cochlear bands equally spaced on an ERB scale from 80-8000 Hz. The average correlation between envelopes before and after vocoding is about 0.9. This result suggests that our 64-channel envelope-vocoding procedure leaves the within-band cochlear-level envelopes largely intact, thereby allowing us to unambiguously attribute vocoding effects to TFS cues.

The stimulus used for online volume adjustment was separately generated, and consisted of running speech mixed with four-talker babble. The speech and babble samples were both obtained from the QuickSIN corpus (Killion et al., 2004), and repeated over time to obtain a total stimulus duration of 20 s (so as to give subjects adequate time to adjust their computer volume). The volume adjustment stimulus was also designed to have a root mean square (RMS) value that was 75% of the dB difference between the softest and loudest stimuli in the study. This ensured that once subjects had adjusted their computer volume, the stimuli used in the main consonant identification tasks were never too loud for subjects, even at adverse SNRs.

2.2 Participants

Data were collected online from anonymous subjects recruited using Prolific.co. The subject pool was restricted using a screening study developed by Mok et al. (2021). The screening study contained three parts: (i) a core survey that was used to restrict subjects based on age (to exclude significant age-related hearing loss because we cannot screen for it in online experiments), whether or not they were native speakers of North American English (because we used North American speech stimuli), history of hearing and neurological diagnoses if any, and whether or not they had persistent tinnitus, (ii) headphone/earphone checks, and (iii) a speech-in-babble-based hearing screening. Subjects who passed the screening study were invited to participate in the consonant identification study, and when they returned, headphone/earphone checks were performed again. These procedures were validated in previous work, where they were shown to successfully select participants for near-normal hearing status, attentive engagement, and stereo headphone use (Mok et al., 2021).

The subjects used in the consonant identification experiments were aged 18–55 years, and self-reported not having any hearing loss, neurological disorders, or persistent tinnitus. All subjects were US/Canada residents, US/Canada born, and native speakers of North American English. In addition, all subjects had completed at least 40 previous studies on Prolific and had $> 90\%$ of them approved (Prolific allows researchers to reject participant submissions if there is clear evidence of non-compliance with instructions or poor attention). Finally, all subjects also passed the headphone/earphone checks and speech-in-babble-based hearing screening (Mok et al., 2021). Subjects provided informed consent in accordance with remote testing protocols approved by the Purdue University Institutional Review Board (IRB).

2.3 Experimental design

We conducted three nearly identical consonant-identification experiments to test for replicability of any main effect of fine structure. The experiments were designed with the goal of contrasting intact and vocoded conditions (i.e., stimuli with and without fine structure), while roving the levels of all other experimental variables (i.e., random effects). Each experiment presented, in random order, one stimulus repetition for each of the 20 consonants across all four talkers and all five experimental conditions. Within a given experiment, in creating each intact or vocoded SiB stimulus, babble instances (i.e., examples or realizations) were randomly chosen from a list comprising 72 different four-talker babble maskers (see Section 2.1); thus, the babble instances that were used for a particular consonant and talker were not the same between the intact and vocoded SiB conditions. To test whether the main effects of fine structure generalized when the babble instances used were varied across experiments, we used a different random pairing of masker instances across consonants, talkers, and conditions in Expt 2 compared to Expt 1. However, Expt 3 used, as a sanity check while testing replication of effects, the exact same stimuli as Expt 2. Thus, the only difference in the stimuli between the experiments was in the particular instance of babble that was paired with a particular consonant, talker, and SiB condition (intact, and vocoded). As observed by Zaar and Dau (2015), when effects are instance-specific, different realizations of the same masker random process can contribute significantly larger variability to consonant identification measurements than the amount of within-listener variability. Thus, our study design of varying babble instances across the three experiments helps to disambiguate any effects of vocoding from masker-instance effects.

We used 25 subjects per talker (subject overlap between talkers was not controlled) in each of the three experiments. With four talkers, this yielded 100 subject-talker pairs, or samples, per experiment. There was no overlap between experiments in the particular set of 100 samples that each used, i.e., samples were independent across experiments. Within each experiment, talker, and condition, all subjects performed the task with the same stimuli. Moreover, all condition effect contrasts were computed on a within-subject basis, and averaged across subjects.

Each of the three experiments had three parts: (i) Headphone/earphone checks, (ii) Demo, and (iii) Test (which was the main stage of the experiment). Each of these three parts had a volume-adjustment task at the beginning. In this task, subjects were asked to make sure that they were in a quiet room and wearing wired (not wireless) headphones or earphones. They were instructed not to use desktop/laptop speakers. They were then asked to set their computer volume to 10–20% of the full volume, following which they were played a speech-in-babble stimulus and asked to adjust their volume up to a comfortable but not too loud level. Once subjects had adjusted their computer volume, they were instructed to not adjust the volume anymore during the experiment, as that could lead to sounds being too loud or soft.

We used the paradigm of Mok et al. (2021) for headphone/earphone checks. In this paradigm, subjects first performed the Woods et al. (2017) task (to verify headphone/earphone use), followed by a second task that involved detection of fluctuating interaural correlation in broadband noise (to specifically test if headphones/earphones were used in both ears). Only those subjects who scored greater than 65% in each of these two tasks were allowed to proceed to the next (Demo) stage of the experiment.

In the Demo part, subjects performed a short training task designed to familiarize them with how each consonant sounds, and with the consonant-identification paradigm. Subjects were instructed that in each trial they would hear a voice say "You will mark *something* please." They were told that at the end of the trial, they would be given a set of options for *something*, and that they would have to click on the corresponding option. Consonants were first presented in quiet, and in sequential order starting with /b/ and ending with /ʒ/. Note that this order was matched in the consonant options shown on the screen at the end of each trial. Thus, after the stimulus ended in each trial, subjects were shown a list of 20 consonants along with example words in which each occurred, and tasked to mark the consonant they heard. The Demo used the same talker's voice as the Test stage of the experiment. After subjects had heard all consonants sequentially in quiet, they were next tasked with identifying consonants presented in random order and spanning the same set of listening conditions as the Test stage. Subjects were instructed to ignore any background noise and only listen to the particular voice saying "You will mark *something* please." Only subjects who scored $\geq 85\%$ in the Demo's Speech in Quiet control condition were selected for the Test stage, so as to ensure that we included only subjects who understood and were able to perform the task.

In the Test stage (i.e., the main part of the experiment), subjects were given similar instructions as in the Demo, but told to expect trials with background noise from the beginning (rather than midway through the task as in the Demo). In both Demo and Test, the background noise (babble or stationary noise) started first for those trials that presented noisy speech and continued for the entire duration of the trial, while the target speech started 1 s after the background started. This was done to help cue the subjects' attention to the stimulus before the target sentence was played. In both Demo and Test, in order to promote engagement with the task, subjects received feedback in every trial as to whether or not their responses were correct. However, subjects were generally not given the correct answer to avoid over-training to the acoustics of how each consonant sounded across the different conditions, except for the first sub-part of the Demo where subjects heard all consonants in quiet in sequential order.

2.4 Hardware

Subjects performed the tasks using their personal computers and headphones/earphones. Our online infrastructure included checks to prevent the use of mobile devices.

2.5 Data preprocessing

Only samples with intelligibility scores $\geq 85\%$ for the Speech in Quiet control condition in the Test stage were included in results reported here. The remaining outlier samples were excluded from further analyses as a data quality control measure.

2.6 Quantifying confusion matrices

The 20 English consonants used in this study were assigned the phonetic features described in Table 1.

Table 1. Phonetic features of the 20 English consonants used in this study.

Consonant	Voicing	Manner of articulation (MOA)	Place of articulation (POA)	Binary POA
/b/	Voiced	Stop	Bilabial	Front
/ʧ/	Unvoiced	Affricative	Palatal	Back
/d/	Voiced	Stop	Alveolar	Back
/ð/	Voiced	Fricative	Dental	Front
/f/	Unvoiced	Fricative	Labiodental	Front
/g/	Voiced	Stop	Velar	Back
/ʤ/	Voiced	Affricative	Palatal	Back
/k/	Unvoiced	Stop	Velar	Back
/l/	Voiced	Liquid	Alveolar	Back
/m/	Voiced	Nasal	Bilabial	Front
/n/	Voiced	Nasal	Alveolar	Back
/p/	Unvoiced	Stop	Bilabial	Front
/r/	Voiced	Liquid	Palatal	Back
/s/	Unvoiced	Fricative	Alveolar	Back
/ʃ/	Unvoiced	Fricative	Palatal	Back
/t/	Unvoiced	Stop	Alveolar	Back
/θ/	Unvoiced	Fricative	Dental	Front
/v/	Voiced	Fricative	Labiodental	Front
/z/	Voiced	Fricative	Alveolar	Back
/ʒ/	Voiced	Fricative	Palatal	Back

The consonant identification data collected in the Test stage of each experiment were used to construct consonant confusion matrices (pooled over samples) separately for each condition. Overall intelligibility was normalized to 60% for intact and vocoded SiB, and to 90% for intact and vocoded SiQuiet by scaling the confusion matrices such that the sum of the diagonal entries was the desired intelligibility. The resulting consonant confusion matrices were used to construct voicing, POA, and MOA confusion matrices by pooling over all consonants. In order to test our hypothesis that voicing, POA, and MOA confusion patterns will be the same for intact and envelope-vocoded speech in babble (after matching intelligibility), we computed the difference between intelligibility-matched intact and vocoded SiB confusion matrices. Confusion matrix differences were then compared with appropriate null distributions of zero differences to extract statistically significant differences (see Section 2.7). A similar procedure was used to test whether fine structure conveys speech content in quiet over and above what is conveyed by envelopes, but by pooling data across all three experiments when constructing confusion matrices for intact and vocoded SiQuiet (versus examining effects separately for each experiment, as was done for intact and vocoded SiB). This data pooling across experiments was performed to improve statistical power because of the relatively high overall intelligibility in quiet (i.e., very few confusions).

2.7 Statistical analysis

To examine the role of fine structure in conveying speech content, we computed the difference in the voicing, POA, and MOA confusion matrices between intact and vocoded conditions, separately for speech in babble and speech in quiet. Permutation testing (Nichols and Holmes, 2002) with multiple-comparisons correction at 5% false-discovery rate (FDR; Benjamini and Hochberg, 1995) was used to extract significant differences in the confusion patterns. The null distributions for permutation testing were obtained using a non-parametric shuffling procedure, which ensured that the data used in the computation of the null distributions had the same statistical properties as the measured confusion data. Separate null distributions were generated for speech in babble and speech in quiet, and for the different phonetic categories. Each realization from each null distribution was obtained by following the same computations used to obtain the actual "intact - vocoded" confusion matrices, but with random shuffling of intact versus vocoded condition labels corresponding to the measurements. This procedure was repeated with 10,000 distinct randomizations for each null distribution.

To quantify the degree to which statistically significant "intact - vocoded" confusion differences were replicated across the three experiments, we used simple Pearson correlation and derived the p-value for the correlation using Fisher's approximation (Fisher, 1921). Although the entries of each difference matrix are not strictly independent (which can cause p-values to be underestimated), we considered this p-value approximation to be adequate given that visual inspection of replication results were unambiguous.

2.8 Signal-detection theoretic analysis

A signal-detection theoretic analysis (Green et al., 1966) was used to calculate the bias, i.e., the shift in the classification boundary, in the average subject's percept of voicing for target speech in babble relative to an unbiased ideal observer (i.e., a classifier that optimally uses the acoustics to arrive at a speech-category decision) (see Fig. 2). The extent to which this bias was altered by vocoding was then quantified. This analysis was motivated by the finding that vocoding had a significant and replicable effect on voicing confusions for speech in babble across the three experiments in our study.

Let us define the null and alternative hypotheses for the voicing categorization performed by listeners. Let $\mathcal{H}0$ be the null hypothesis that an unvoiced consonant was presented, and let $\mathcal{H}1$ be the alternative hypothesis that a voiced consonant was presented. Let FA be the probability of a false alarm, and HR be the hit rate. The FA and HR values for each experiment and condition were directly obtained from the voicing confusion matrix (pooled over samples and consonants) corresponding to that experiment and condition.

The cutoff C (or decision boundary) for the average subject's perceptual decision on whether or not to reject $\mathcal{H}0$, d', and listener bias B (expressed as a percentage relative to an unbiased ideal observer's cutoff) were calculated separately for each experiment and condition (intact versus vocoded SiB) as:

$$C = \phi(1 - FA),$$

$$d' = \phi(1 - FA) - \phi(1 - HR),$$

and

$$B = \frac{(C - d'/2) \times 100}{d'/2},$$

where ϕ is the inverse of the standard normal cumulative distribution.

The change in the listener bias between the intact and vocoded SiB conditions was derived as:

$$B_{vocoded} - B_{intact},$$

where $B_{vocoded}$ and B_{intact} are the biases in the vocoded and intact SiB conditions, respectively.

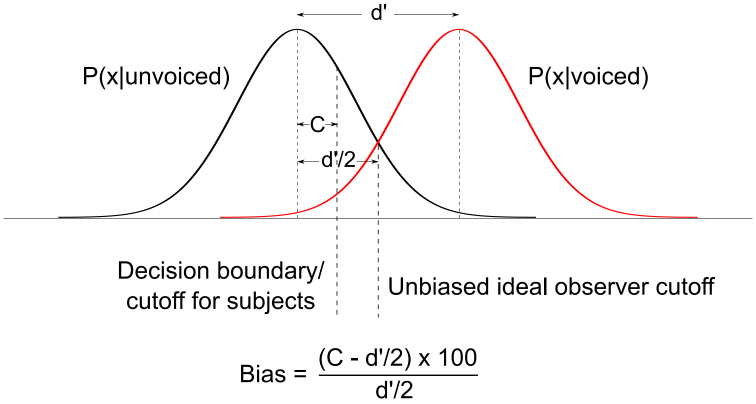

Figure 2. Illustration of a decision-theoretic quantification of speech categorization bias. x denotes the internal decision variable. Bias is quantified as the percent shift in the average listener's cutoff (or decision boundary) relative to an unbiased ideal observer's cutoff. The cutoff values for the average listener and the ideal observer are estimated from the false-alarm and hit rates in the data.

2.9 Software accessibility

Subjects were directed from Prolific to the SNAPlab online psychoacoustics infrastructure (https://snaplabonline.com; Mok et al., 2021) to perform the study. Offline data analyses were performed using custom software in Python (Python Software Foundation, https://www.python.org) and MATLAB (The MathWorks, Inc., Natick, MA). Copies of all custom code can be obtained from the authors.

3 Results

Figure 3 shows intelligibility measurements across all conditions and experiments in our study. Approximately equal overall intelligibility was achieved for intact and vocoded SiB due to our careful choice of SNRs for these conditions, based on extensive offline piloting. This allowed small differences in intelligibility to be normalized without loss of statistical power. Overall intelligibility was normalized to 60% for intact and vocoded SiB, and to 90% for intact and vocoded SiQuiet, respectively (as described in Section 2.6), before examining the effects of vocoding on voicing, POA, and MOA confusion patterns.

Given that our data were collected online, we conducted a few different data quality checks. We first examined whether subjects randomly chose a different consonant from what was presented when they made an error, or if there was more structure in the data. Figure 4 suggests that the error patterns in our data have a non-random structure, which supports the validity of our online-collected data.

To further test data quality, we compared consonant confusions for the SiSSN condition with previous lab-based findings, since speech-shaped stationary noise is a commonly used masker in the phoneme confusions literature. Figure 5 shows consonant groups and confusion clusters identified from our SiSSN data. These results closely replicate previous lab-based findings by Phatak and Allen (2007), lending further support to the validity of our online-collected data.

After verifying data quality, we tested our hypothesis that confusion patterns will be the same for intelligibility-matched intact and envelope-vocoded speech in babble. Figure 6 shows the results for voicing confusions. We find that vocoding alters voicing percept for speech in babble by changing listener bias relative to an ideal observer. In particular, there is a greater tendency in

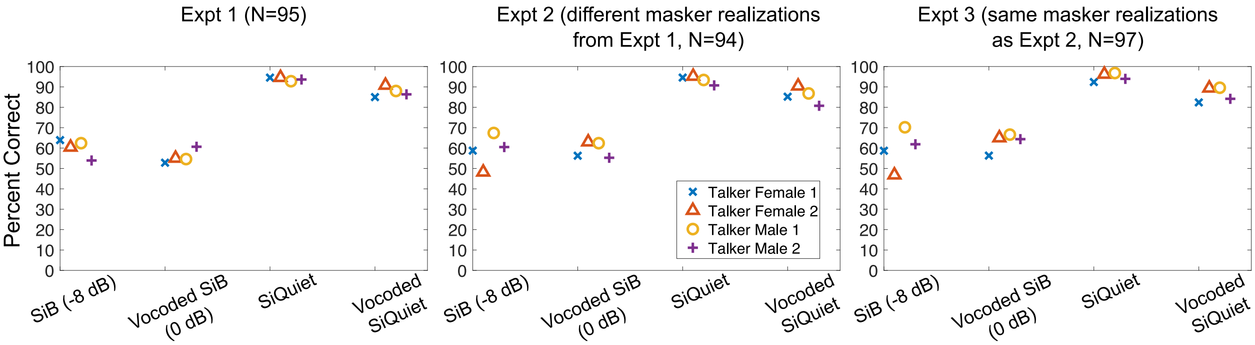

Figure 3. Overall intelligibility was matched between intact and vocoded conditions before comparing confusion patterns. Extensive piloting was done to select SNRs to achieve approximately equal overall intelligibility for intact and vocoded SiB. The actual Prolific data for Expts 1–3 (shown) reflect this, allowing for small differences in intelligibility between intact and vocoded conditions to be normalized without loss of statistical power. Overall intelligibility was normalized to 60% for intact and vocoded SiB, and to 90% for intact and vocoded SiQuiet, respectively (as described in Section 2.6).

the vocoded (versus intact) condition to be biased towards reporting an unvoiced consonant as being heard, despite envelope and place cues being largely preserved. This result suggests that TFS conveys voicing content over what is conveyed by envelopes. This effect is replicated across all three experiments, suggesting that it generalizes across babble instances.

We do not find evidence that TFS conveys either POA or MOA for target speech in babble (Figs. 7 and 8). Although we found significant differences in the confusion patterns between intact and vocoded SiB, effects are replicated only if babble instances are kept constant (between Expts 2 and 3) and not varied (between Expt 1 and Expts 2,3); this suggests that any effects of fine structure on POA or MOA are weak when compared to the babble-instance effect, and do not generalize across instances.

To test whether TFS conveys speech content over and above what is conveyed by envelopes in the quiet condition, we examined the effect of vocoding on consonant confusions in quiet. The results, shown in Figure 9, indicate no significant effects of vocoding on either voicing, POA, or MOA confusions in quiet.

For completeness, the raw confusion matrices for all conditions and experiments are included in Figure 10.

4 Discussion

In the present study, we examined the influence of TFS on consonant confusion patterns by degrading TFS using high-resolution vocoding while carefully controlling intelligibility to match with intact stimuli. Contrary to our hypothesis, we find that TFS conveys voicing information for target speech in babble over and above what is conveyed by envelopes, a result that generalizes across varying babble instances. However, the current study did not find any significant vocoding effects on consonant confusions in quiet even after pooling data across all experiments; instead, overall intelligibility for vocoded SiQuiet was $\sim 90\%$.

The result that TFS conveys voicing for target speech in babble over and above what is conveyed by envelopes is previously unreported to the best of our knowledge. This result deviates from the commonly held view that envelopes convey most speech content (Shannon et al., 1995; Smith et al., 2002). In general, several acoustic cues have been implicated in the categorization of consonant voicing, such as voice onset time (VOT), fundamental frequency at the onset of voicing (onset F0),

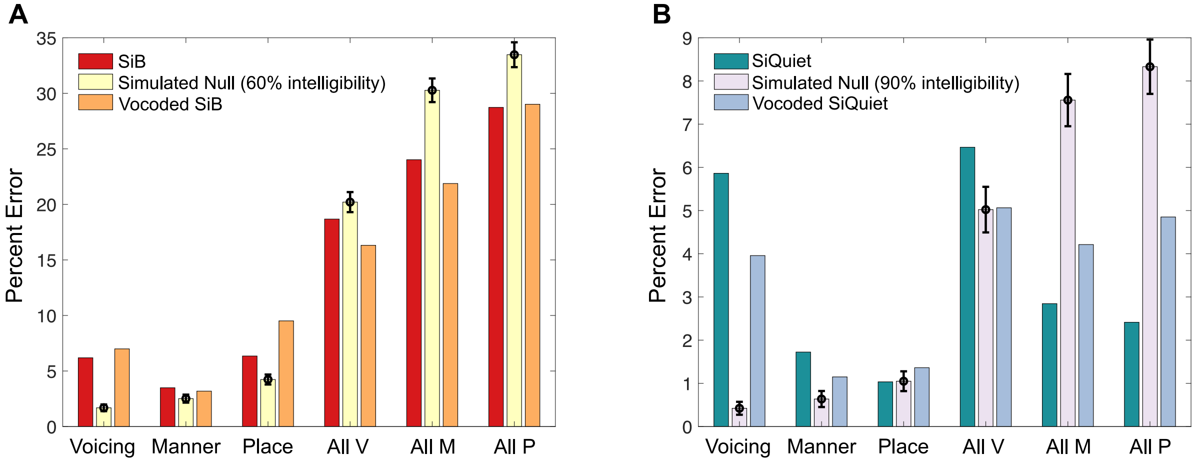

Figure 4. Percent errors (mean and STD from Expt 1) by phonetic category for intact and vocoded SiB (Panel A), and intact and vocoded SiQuiet (Panel B). The labels Voicing, Manner, and Place correspond to when the consonant heard differed from the consonant presented only in voicing, manner of articulation (MOA), or place of articulation (POA), respectively. On the other hand, All V, All M, and All P correspond to when the consonant heard differed from the consonant presented in at least voicing, MOA, or POA, respectively. The expected distribution of errors under the null hypothesis of random confusions was generated separately for Panels A and B, and with 1000 realizations each. Each realization of each null distribution was produced by generating a Bernouilli trial with "success" probability = 60% for Panel A, or 90% for Panel B, followed by uniform-random selection of a different consonant from what was presented if the trial outcome was "failure". Percent errors in our data fall outside the distributions expected from random confusions. This result suggests that the error patterns in our data have a non-random structure, which supports the validity of our online-collected data.

the degree of delay in the onset of the first formant, and the relative amplitude of any aspiration noise in the period between the burst release and the onset of voicing (Francis et al., 2008). Of these, VOT appears to be the dominant cue in quiet (Francis et al., 2008). However, listeners shift reliance to onset F0 when VOT is ambiguous in the presence of noise (Winn et al., 2013; Holt et al., 2018). Our result that vocoding alters voicing percept in noise, but not quiet, is consistent with this result from the cue-weighting literature, and can be attributed to impaired F0 cues resulting from TFS degradation in the vocoded (versus intact) SiB condition. Indeed, voiced sounds (unlike unvoiced) have quasi-periodic acoustic energy reflecting the quasi-periodic vibrations of the vocal folds; this periodicity has a fundamental frequency (F0) that is perceived as pitch (Rosen, 1992). Our finding that TFS can convey additional voicing information beyond envelopes is consistent with the view that the pitch of complex sounds (with resolved harmonics) is coded either via TFS (Meddis and O'Mard, 1997; Smith et al., 2002), or a combination of TFS and tonotopic place (Shamma and Klein, 2000; Oxenham et al., 2004). Indeed, psychophysical studies have found that melody perception (Moore and Rosen, 1979) and fundamental-frequency discrimination (Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2006) are both stronger when conveyed by low-frequency resolved harmonics where the auditory nerve can robustly phase lock to the TFS (Johnson, 1980; Verschooten et al., 2015). Our results from directly manipulating TFS cues also corroborate previous correlational work relating model auditory-nerve TFS coding and voicing reception in noise (Swaminathan and Heinz, 2012).

In the current study, we also find a strong babble-instance effect on POA and MOA confusion patterns. The effects of vocoding on these confusion patterns were not replicated when babble instances were varied between Expts 1 and 2, but were replicated when instances were fixed across Expts 2 and 3. The differences in confusion patterns across varying babble instances can be explained by the relatively short duration of each consonant, which likely leads to small variations

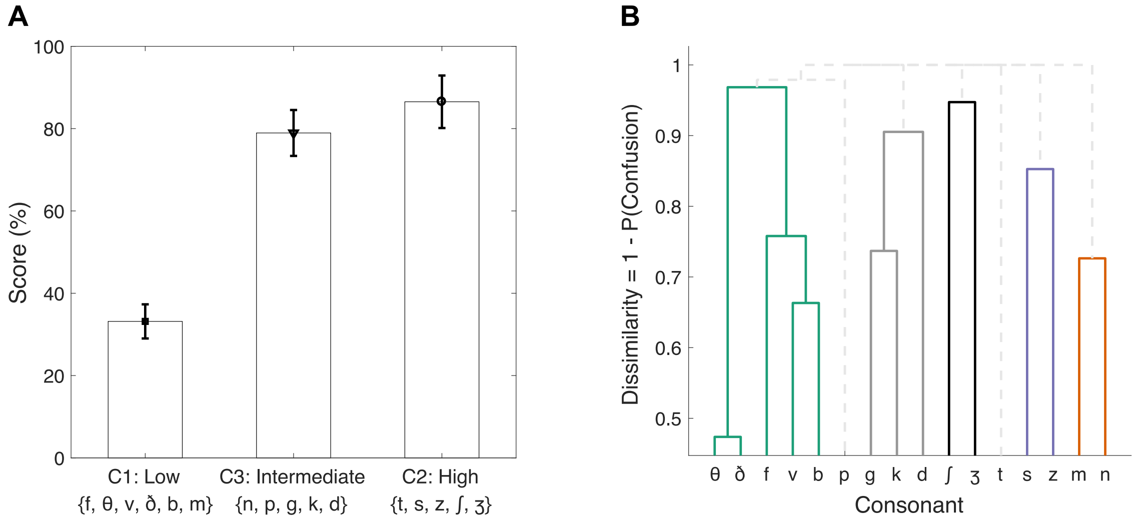

Figure 5. Consonant groups (Panel A) and confusion clusters (Panel B) identified in Expt 1 for speech in speech-shaped stationary noise (SiSSN). Phatak and Allen (2007) found that for a given overall intelligibility, recognition scores vary across consonants. They identified three groups of consonants, "C1", "C3", and "C2" with low, intermediate, and high recognition scores, respectively in speech-shaped noise. Our online-collected data for SiSSN (Panel A) closely replicate that key trend for the groups they identified, after matching the SNR they used. Moreover, using a hierarchical clustering analysis (Ward Jr, 1963) of the consonant confusion matrix (pooled over samples) for SiSSN, we identified perceptual "clusters", i.e., sets where one consonant is confused most with another in the same set (shown as a dendrogram plot in Panel B). Clusters with > 3% probability of confusion share a color. For example, /θ/ and /ð/ form a cluster because they are more confused with each other than with the other consonants; moreover, while /θ/ and /ð/ are less confused with the cluster comprising /f/, /v/, and /b/ than with each other, they are even less confused with all the remaining consonants. The clusters identified here closely replicate the lab-based clustering results of Phatak and Allen (2007), further supporting the validity of our online data.

in the spectral profile of modulation masking across babble instances, despite the average masker modulation spectrum being kept constant (babble is a relatively low-modulation-frequency masker on average; Viswanathan et al., 2021). This interpretation should be further examined in future studies, perhaps using computational modeling to predict instance effects on consonant confusions from variations in modulation masking across short masker instances. Indeed, psychoacoustic literature on speech-in-noise perception (Bacon and Grantham, 1989; Stone and Moore, 2014), neurophysiological studies using EEG (Viswanathan et al., 2021), and and the success of current speech intelligibility models (Dubbelboer and Houtgast, 2008; Relaño-Iborra et al., 2016) show that modulation masking (i.e., masking of the internal representation of temporal modulations in the target by distracting fluctuations from the background) is a key contributor to speech perception in noise.

The fact that we did not find any significant vocoding effects on consonant confusions in quiet even after pooling data across experiments is consistent both with previous behavioral studies that suggested that speech content in quiet is mostly conveyed by envelopes (Shannon et al., 1995; Smith et al., 2002; Elliott and Theunissen, 2009), and with the success of envelope-based cochlear implants in quiet backgrounds (Wilson and Dorman, 2008). However, our finding that voicing cues are degraded in vocoded (versus intact) SiB has implications for current cochlear implants that do not convey TFS cues (Moore, 2008), because babble is a masker that is ubiquitous in everyday listening environments. Indeed, multi-talker babble, which has modulations spanning the range of modulations in target speech, is a more ecological speech-like masker than either stationary noise (which has predominantly high-, but not low-frequency modulations as are present in speech) or

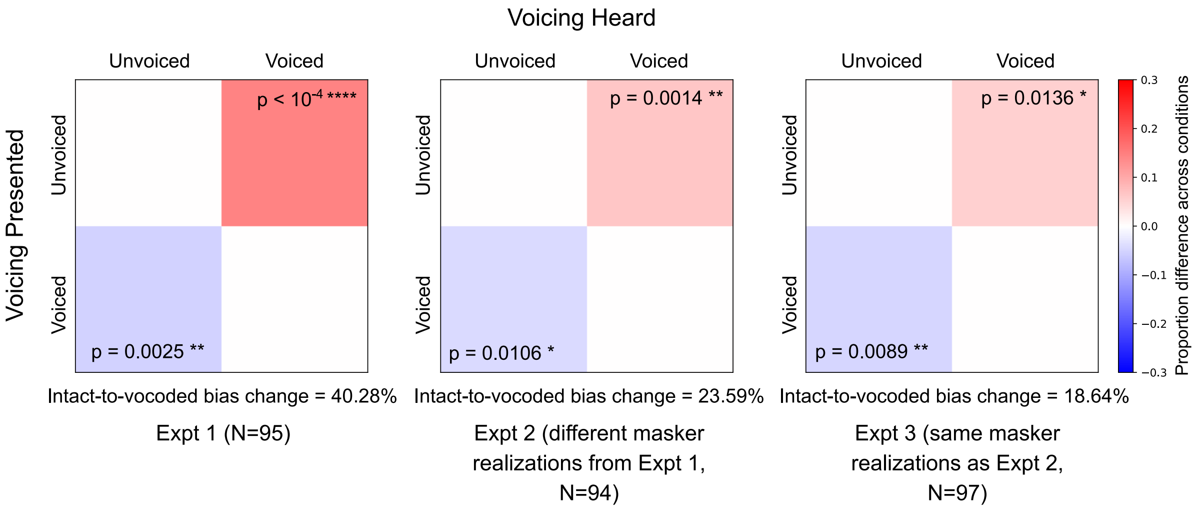

Figure 6. Vocoding alters voicing percept (replicated across the three experiments). Shown are voicing confusion matrix differences (pooled over consonants and samples) between intact and vocoded SiB conditions (SiB - Vocoded SiB), after normalizing overall intelligibility to 60% for each condition. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR). Uncorrected p-values are also indicated for the individual matrix entries. There is a greater tendency in the vocoded (versus intact) condition to be biased towards reporting an unvoiced consonant as being heard, despite envelope and place cues being largely preserved. A detection-theoretic analysis (see Section 2.8) was used to quantify the decision boundary for the average subject's perceptual decision on whether or not to reject the null hypothesis that an unvoiced consonant was presented. The bias or shift in this boundary relative to an unbiased ideal observer was then quantified, and compared between intact and vocoded conditions. Intact-to-vocoded bias changes were found to be about 40%, 24%, and 19% in Expts 1, 2, and 3, respectively. Thus, the result that vocoding biases voicing percept towards unvoiced consonants is replicated across Expts 1–3, which suggests that it generalizes across different babble instances. This finding suggests that fine structure can convey voicing content over and above what is conveyed by envelopes.

even narrow-band syllabic-range AM modulations imposed on stationary noise (Viswanathan et al., 2021), as were used in previous studies (Gnansia et al., 2009; Swaminathan and Heinz, 2012; Winn et al., 2013; Holt et al., 2018). In addition to our finding here that TFS can convey important voicing cues, there is evidence from previous studies that TFS can also aid in source segregation and selective attention (Darwin, 1997; Oxenham and Simonson, 2009; Shinn-Cunningham, 2008; Micheyl and Oxenham, 2010), which can lead to stronger representation of target-speech envelopes in the brain that predicts intelligibility (Viswanathan et al., 2021). The effect of TFS on segregation is in fact reflected in the present study too, where we had to increase the SNR for vocoded SiB by 8 dB relative to intact SiB in order to match their respective overall intelligibility values. Taken together, these results suggest that patients with cochlear implants may benefit from the inclusion of TFS cues for speech recognition in everyday listening environments with multiple talkers or sound sources. This finding should be further examined in future studies using clinical populations.

5 Conclusion

We find evidence that fine structure conveys voicing content for target speech in babble over and above the content conveyed by envelopes, and after controlling for overall performance. This result was robustly replicated when babble instances were varied across independent experiments. Given that babble is a masker that is ubiquitous in everyday environments, this finding has implications for assistive listening devices such as cochlear implants that do not currently provide TFS cues.

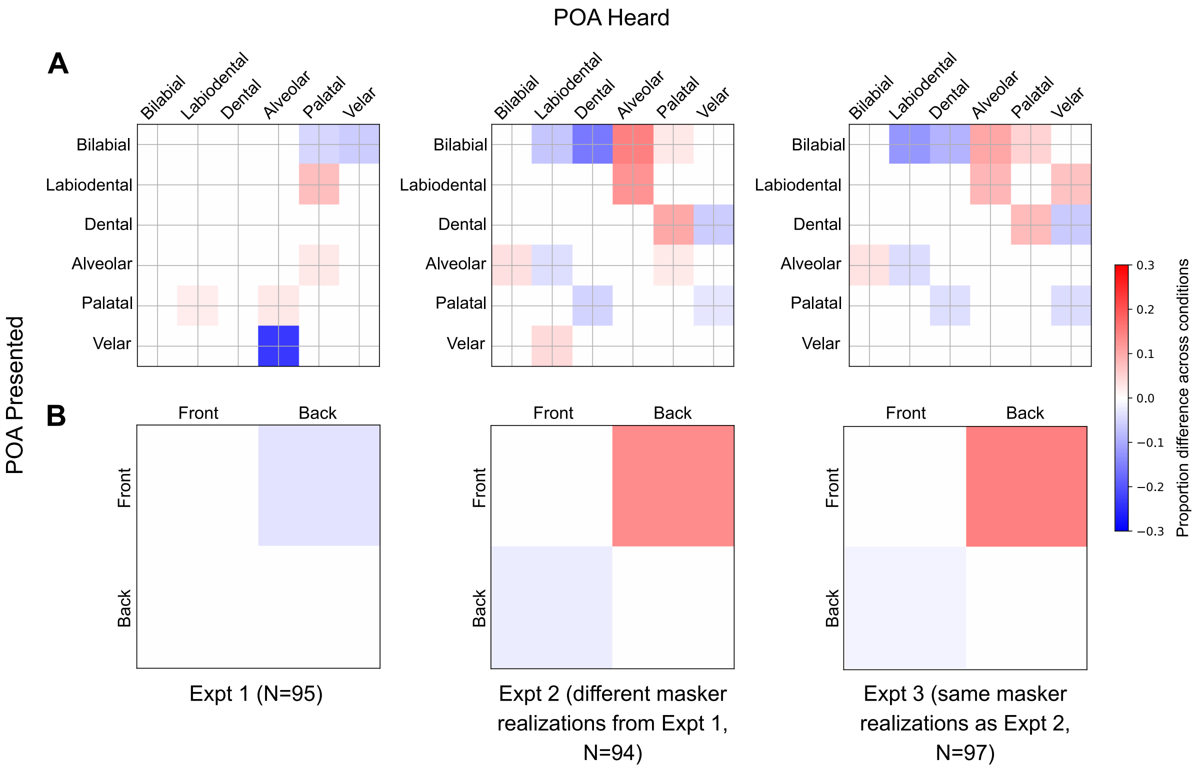

Figure 7. Effects of vocoding on POA are not replicated across varying babble instances. Shown are POA confusion matrix differences (pooled over consonants and samples) between intact and vocoded SiB (SiB - Vocoded SiB), after normalizing overall intelligibility to 60% for each condition. Panel A shows full (5x5) matrices, whereas Panel B shows simplified (binary) matrices after collapsing over front versus back places of articulation. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR). Results are not replicated when varying babble instances (between Expts 1 and 2; $R^2 = 2 \times 10^{-6}$, $p = 0.99$), but are replicated when stimuli are kept constant (between Expts 2 and 3; $R^2 = 0.85$, $p = 3.77 \times 10^{-13}$). The fact that results do not generalize across babble instances suggests a greater babble-instance effect than any effects due to manipulating fine structure.

6 Acknowledgments

This research was supported by grants from the National Institutes of Health [F31DC017381 (to V.V.), R01DC009838 (to M.G.H.), and R01DC015988 (to B.G.S.-C.)]. We thank Hari Bharadwaj for access to online psychoacoustics infrastructure (https://snaplabonline.com; Mok et al., 2021). We also thank Agudemu Borjigan, Andrew Sivaprakasam, François Deloche, Hari Bharadwaj, Ivy Schweinzger, Ravinderjit Singh, and Satyabrata Parida for valuable feedback on an earlier version of this manuscript.

7 Author Contributions

V.V., B.G.S.-C., and M.G.H. designed research; V.V. performed research; V.V. analyzed data; V.V. wrote the paper with edits from B.G.S.-C. and M.G.H.

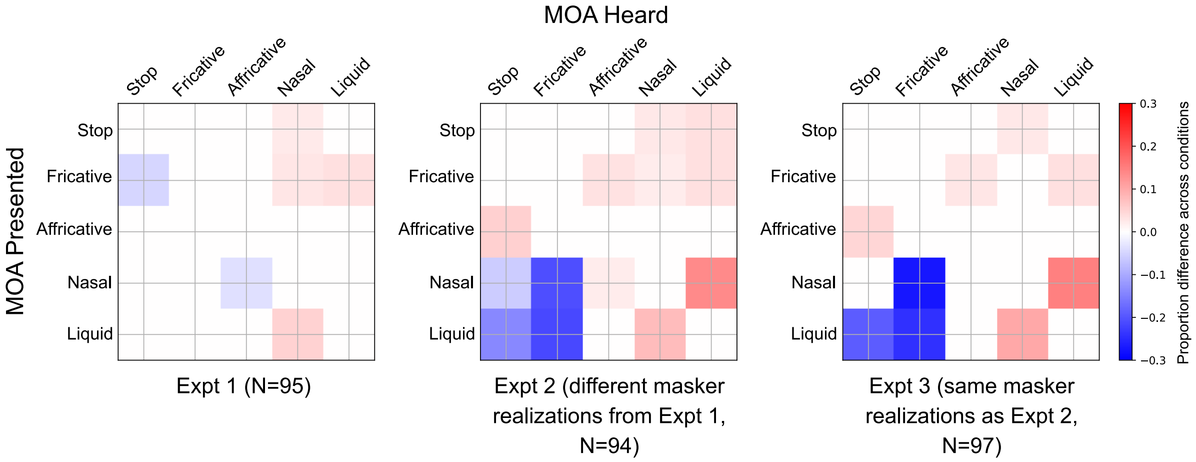

Figure 8. Effects of vocoding on MOA are not replicated across varying babble instances. Shown are MOA confusion matrix differences (pooled over consonants and samples) between intact and vocoded SiB (SiB - Vocoded SiB), after normalizing overall intelligibility to 60% for each condition. Only significant differences are shown, after permutation testing with multiple-comparisons correction (5% FDR). Results are not replicated when varying babble instances (between Expts 1 and 2; $R^2 = 0.03$, $p = 0.44$), but are replicated when stimuli are kept constant (between Expts 2 and 3; $R^2 = 0.94$, $p = 1.44 \times 10^{-12}$). Since results do not generalize across babble instances, it can be inferred that there is a stronger babble-instance effect than any effects due to fine structure.

References

Bacon, S. P. and Grantham, D. W. (1989). Modulation masking: Effects of modulation frequency, depth, and phase. *J Acoust Soc Am*, 85(6):2575–2580.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Series B Stat Methodol*, pages 289–300.

Bernstein, J. G. and Oxenham, A. J. (2006). The relationship between frequency selectivity and pitch discrimination: Effects of stimulus level. *J Acoust Soc Am*, 120(6):3916–3928.

Darwin, C. J. (1997). Auditory grouping. *Trends Cogn Sci*, 1(9):327–333.

Dubbelboer, F. and Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *J Acoust Soc Am*, 124(6):3937–3946.

Dubno, J. R. and Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *J Acoust Soc Am*, 69(1):249–261.

Elliott, T. M. and Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Comput Biol*, 5(3):e1000302.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, 1:1–32.

Francis, A. L., Kaganovich, N., and Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in english. *J Acoust Soc Am*, 124(2):1234–1251.

Gilbert, G. and Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. *J Acoust Soc Am*, 119(4):2438–2444.

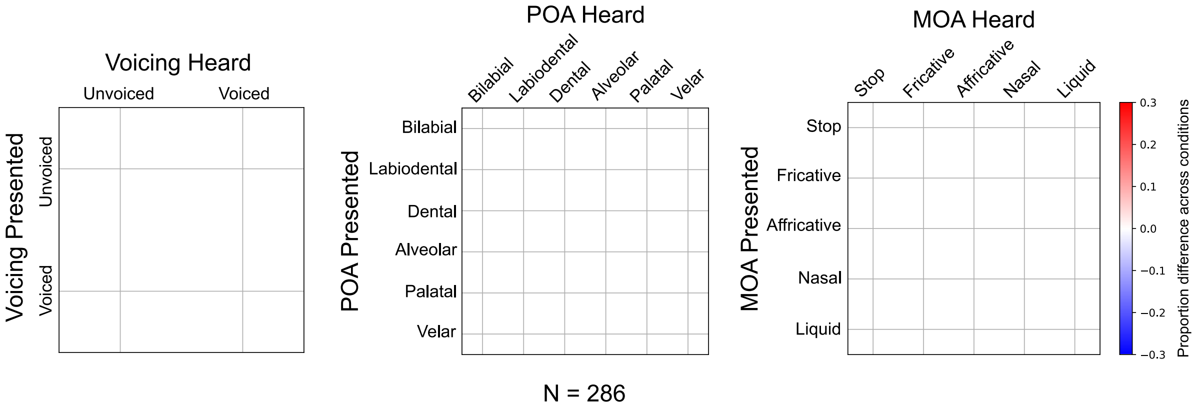

Figure 9. Vocoding does not significantly alter confusions in quiet. Shown are voicing, POA, and MOA confusion matrix differences (pooled across all experiments, consonants, and samples) between intact and vocoded speech in quiet (intact - vocoded SiQuiet), after normalizing overall intelligibility to 90% for each condition. No significant differences are found (as shown), after permutation testing with multiple-comparisons correction (5% FDR). Thus, we do not find any significant effects of degrading fine structure on the perception of either voicing, POA, or MOA in quiet.

Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear Res*, 47(1):103–138.

Gnansia, D., Péan, V., Meyer, B., and Lorenzi, C. (2009). Effects of spectral smearing and temporal fine structure degradation on speech masking release. *J Acoust Soc Am*, 125(6):4023–4033.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics*, volume 1. Wiley New York.

Hilbert, D. (1906). Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen. Vierte Mitteilung. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1906:157–228.

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., and Dick, F. (2018). Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing. *Hear Res*, 366:50–64.

Hopkins, K. and Moore, B. (2010). The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *J Acoust Soc Am*, 127(3):1595–1608.

Houtsma, A. J. and Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *J Acoust Soc Am*, 87(1):304–310.

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J Acoust Soc Am*, 68(4):1115–1122.

Joris, P. X. and Yin, T. C. (1992). Responses to amplitude-modulated tones in the auditory nerve of the cat. *J Acoust Soc Am*, 91(1):215–232.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., and Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*, 116(4):2395–2405.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc Natl Acad Sci U S A*, 103(49):18866–18869.

Meddis, R. and O'Mard, L. (1997). A unitary model of pitch perception. *J Acoust Soc Am*, 102(3):1811–1820.

Micheyl, C. and Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hear Res*, 266(1-2):36–51.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *J Acoust Soc Am*, 27(2):338–352.

Mok, B. A., Viswanathan, V., Borjigin, A., Singh, R., Kafi, H. I., and Bharadwaj, H. M. (2021). Web-based psychoacoustics: Hearing screening, infrastructure, and validation. *bioRxiv*, page DOI: 10.1101/2021.05.10.443520.

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *J Assoc Res Otolaryngol*, 9(4):399–406.

Moore, B. C. and Rosen, S. M. (1979). Tune recognition with reduced pitch and interval information. *Q J Exp Psychol*, 31(2):229–240.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*, 15(1):1–25.

Oxenham, A. J., Bernstein, J. G., and Penagos, H. (2004). Correct tonotopic representation is necessary for complex pitch perception. *Proc Natl Acad Sci USA*, 101(5):1421–1425.

Oxenham, A. J. and Simonson, A. M. (2009). Masking release for low-and high-pass-filtered speech in the presence of noise and single-talker interference. *J Acoust Soc Am*, 125(1):457–468.

Phatak, S. A. and Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *J Acoust Soc Am*, 121(4):2312–2326.

Qin, M. and Oxenham, A. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J Acoust Soc Am*, 114(1):446–454.

Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., and Dau, T. (2016). Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain. *J Acoust Soc Am*, 140(4):2670–2679.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci*, 336(1278):367–373.

Shamma, S. and Klein, D. (2000). The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. *J Acoust Soc Am*, 107(5):2631–2644.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304.

Sheft, S., Ardoint, M., and Lorenzi, C. (2008). Speech identification based on temporal fine structure cues. *J Acoust Soc Am*, 124(1):562–575.

Shinn-Cunningham, B. (2008). Object-based auditory and visual attention. *Trends Cogn Sci*, 12(5):182–186.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90.

Stone, M. A. and Moore, B. C. (2014). On the near non-existence of “pure” energetic masking release for speech. *J Acoust Soc Am*, 135(4):1967–1977.

Swaminathan, J. and Heinz, M. G. (2012). Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise. *J Neurosci*, 32(5):1747–1756.

Verschooten, E., Robles, L., and Joris, P. X. (2015). Assessment of the limits of neural phase-locking using mass potentials. *J Neurosci*, 35(5):2255–2268.

Viswanathan, V., Bharadwaj, H. M., Shinn-Cunningham, B. G., and Heinz, M. G. (2021). Modulation masking and fine structure shape neural envelope coding to predict speech intelligibility across diverse listening conditions. *bioRxiv*, page DOI: 10.1101/2021.03.26.437273.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*, 58(301):236–244.

Wilson, B. S. and Dorman, M. F. (2008). Cochlear implants: a remarkable past and a brilliant future. *Hear Res*, 242(1-2):3–21.

Winn, M. B., Chatterjee, M., and Idsardia, W. J. (2013). Roles of voice onset time and f0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering. *J Speech Lang Hear Res*, 56:1097–1107.

Woods, K. J., Siegel, M. H., Traer, J., and McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Atten Percept Psychophys*, 79(7):2064–2072.

Zaar, J. and Dau, T. (2015). Sources of variability in consonant perception of normal-hearing listeners. *J Acoust Soc Am*, 138(3):1253–1267.

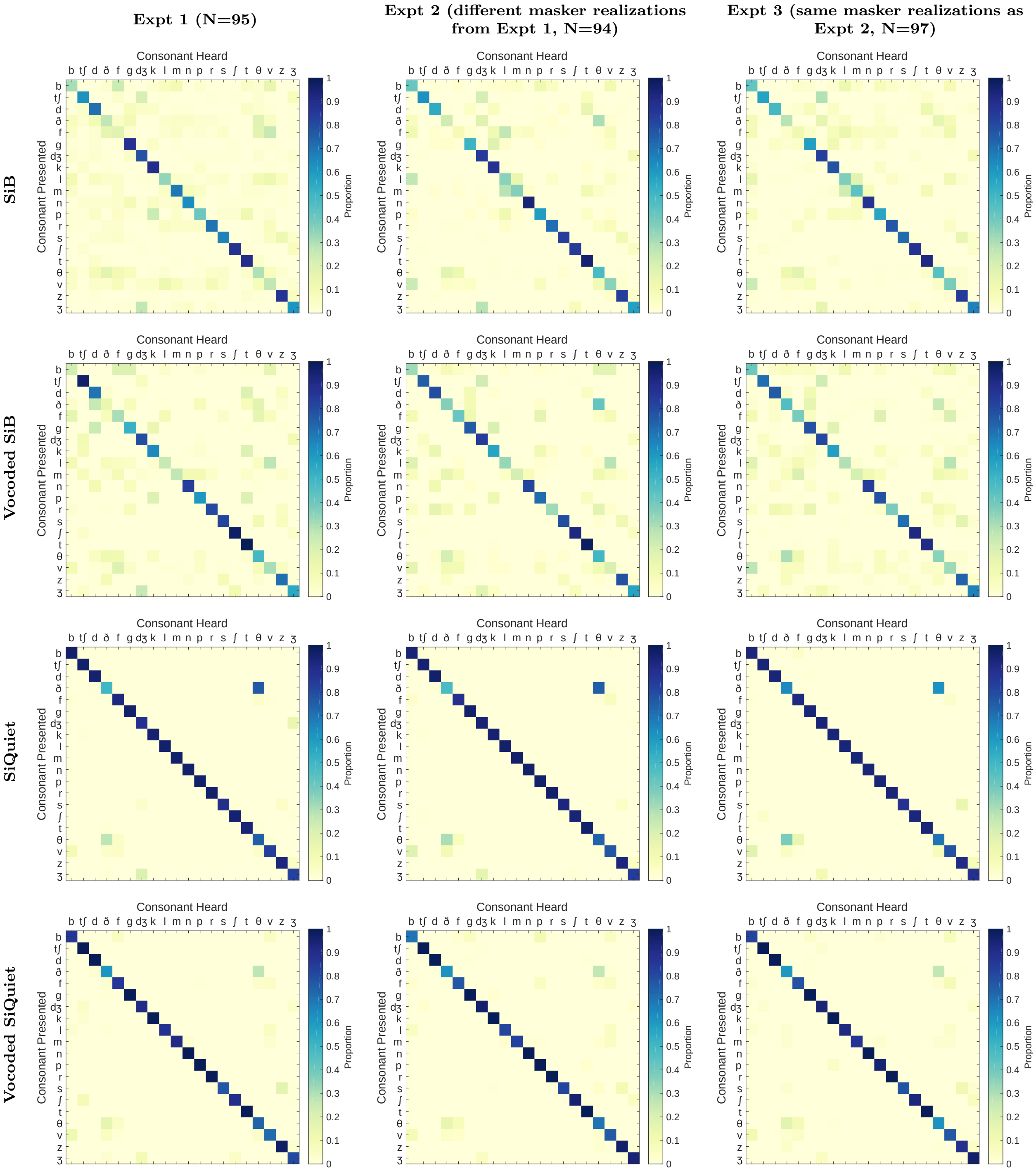

Figure 10. Raw confusion matrices for all conditions and experiments (pooled over samples). Note that overall intelligibility is 60% for the SiB and vocoded SiB conditions, and 90% for the SiQuiet and vocoded SiQuiet conditions, respectively.